DOCUMENT RESUME

ED 441 789 SP 039 247

AUTHOR Hall, Lisa A.; McDonald, Jo-Anne L.
TITLE Measuring Change in Teachers' Perceptions of the Impact that Staff Development Has on Teaching.
PUB DATE 2000-04-00
NOTE 20p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 24-28, 2000).
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Attitude Change; Attitude Measures; Elementary Education; Elementary School Teachers; *Evaluation Methods; *Faculty Development; Inservice Teacher Education; Program Evaluation; Reading Instruction; *Teacher Attitudes; Teacher Improvement

ABSTRACT

This study compared data from an evaluation of one school district's teacher staff development programs, using the True Score Theory and Item Response theory. Participants were elementary school teachers who reported on 20 staff development programs in reading. They completed the Teachers' Perception of the Impact of a Staff Development (TPISD), which examined how they thought, felt, and taught following staff development. The TPISD was given during the spring and the fall to assess whether the initial reported impact changed once teachers applied what they learned in the classroom. Data were analyzed under the assumptions of the True Score Theory and the Item Response theory. Use of the True Score Theory for evaluating change in scores on the TPISD across 4 months provided no evidence of change. Item Response Theory analysis indicated that 26 percent of the scores changed significantly across 4 months, which is far more than would be expected by chance using a 95-percent confidence interval. By using corrected person scores obtained during the stability analysis, more certainty was gained that change in scores was due to changes in the level of the measured variable rather than changes in the measure itself. (Contains 11 references.) (SM)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 441 789

Measuring Change in Teachers' Perceptions of the Impact that Staff Development has on Teaching

Lisa A. Hall, Ph.D.
University of Denver
Boulder Valley School District

Jo-Anne L. McDonald
University of Denver

Paper presented at the Annual Meeting
American Educational Research Association
New Orleans, Louisiana
April, 2000

BEST COPY AVAILABLE

SP039247

L. Hall

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☐ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Change Analysis of Person Scores Over Time

Measurement of change presents a "nasty challenge" (Wright, 1996, pg. 478). The challenge is to measure persons and items in the same clearly defined frame of reference at both time points, so that measurements of change will have unambiguous meaning. Though program evaluators may be examining the change in persons from time 1 to time 2, the functioning of test items and rating scales may also have changed. Only if the items are invariant from group to group and from time to time can meaningful comparisons of person scores be made (Wright & Masters, 1982).

Traditionally, summed scores from two administrations of a measure given to the same persons are compared and the difference between scores is attributed to changes in the latent trait. This posttest score minus the pretest score is called a gain or difference score (Gall, Borg, & Gall, 1996). There are several problems with the interpretation of gain scores though not all researchers agree to what extent these difficulties should limit their use (Collins, 1996; Williams & Zimmerman, 1996). These problems include the assumption of equal intervals and inconsistent interpretation of items or response options.

The equal interval assumption relates to a measurement scale formed by raw scores which is assumed to be acting as a linear measurement system (Linacre, 1998, April). Equal intervals are believed to exist between all points on a test, yet this assumption is almost never valid for educational or psychological measures (Gall et al., 1996). With the use of Item Response Theory (IRT) models in the development of a measure, the assumption of equal intervals can be met (Wright & Masters, 1982). IRT models involve the placing of items and persons on a common, equal-interval scale. This results in linear

measures which can be analyzed using traditional statistics and which allow for the person level analysis of change in pre/post test scores.

The second noted problem with the traditional analysis of gain scores is that it does not take into account the possibility that respondents may interpret the items or the rating scale options differently on the two occasions (Wright, 1996). Item Response Theory (IRT) models are able to address this issue because they contain one or more parameters for each item and person, with these parameters being invariant. The major advantage of invariance is that person parameters are not test-dependent and item parameters are not sample-dependent. This means that similar estimates of person ability will be derived regardless of which items are completed, and that similar item parameters will be derived regardless of the ability or latent trait level inherent in the persons taking the measure. Thus, invariance allows for predictions about how a person with a certain level of a trait will respond to an item with a certain level of difficulty. With these predictions, one can also assess whether persons responded to items in the expected pattern on the same measure given at two different time points. This comparison of obtained patterns with predicted patterns allows changes in scores to be partialed out into changes due to an intervention and changes due to the measurement instrument itself. If observed patterns of responses fit the expected pattern of responses over the two administrations, then change can be attributed to change in the latent trait. If observed patterns of responses differ from expected patterns, then a change in the instrument functioning is supported.

Despite the dramatic increase in the use of IRT, a survey of the literature on the evaluation of staff developments found no reference to IRT in the development of measures used in the evaluations. This could be due to IRT being mathematically

complex when compared to true score theory, or that many researchers are unfamiliar with current models in item response theory which in turn limits their use. Whatever the reasons, a comparison of findings using both True Score theory and Item Response theory could be useful in demonstrating to program evaluators the advantages and limitations of both theories when evaluating change in persons as a result of an intervention. The current study used data gathered from an evaluation of a district's teacher staff development programs in order to provide such a comparison of analyses.

Method

Participants

The school district involved in this study is located in a suburb of a large mid-western city. Potential participants were teachers who had completed a reading staff development program through the district during the summer and fall of the 1998-199 school year. The resulting sample of 166 teachers was drawn from all of the 29 elementary schools located within the participating school district. Teaching assignments covered the range of known teaching assignments within the district including grade level teachers (n = 80), split grade level or split assignment teachers (n = 29), reading recovery teachers (n = 19), reading teachers (n = 10), and special education teachers including gifted/talented and ESL (n = 14). The remainder of the teachers (n = 14) either did not report the information or could not be placed in one of the above categories. The reported mean years of full-time teaching experience was 13.61 (SD = 8.02), with the range being 1 to 37 years. The most common reported level of education was the M.A.+ category (56%), with a B.A.+ (33.7%), M.A. (4.8%), B.A. (1.8%), and Ph.D. (n = 1)

following in descending frequency. The sample for administration two consisted of 162 teachers from the pool of 166 teachers who completed the first TPISD. Four teachers were dropped from the original pool because three of the surveys could not be traced to an identification number, and the fourth survey was completed by the person dropped during an initial Rasch analysis. The number of surveys returned for the second administration was 152 of 162, for a return rate of 94%. Informal evaluation of the demographics for those not returning surveys revealed no pattern of differential dropout.

Staff Developments

The sample of teachers reported on a total of 20 different staff developments, all in the area of reading. Initially is was expected that teachers would report only on the 12 staff developments run through the district. However, the instructions stated for teachers to report on a staff development on reading they had taken during the summer and fall, and therefore, teachers also reported on 8 additional staff developments offered through their schools. As is noted in the results section, this did pose not a problem for the analyses using IRT, but was a problem in the True Score theory analyses.

Instrument

The Teachers' Perception of the Impact of a Staff Development (TPISD) is a 25-item rating scale measure developed in order to provide a teachers' perspective in an evaluation of staff development programs (Appendix A). The measure includes items related to expected changes in the way a teacher thinks, feels, and teaches after having participated in a staff development program. The TPISD was given at two different time periods, fall and spring, in order to access if the initial impact reported by a teacher

changed once they had had the opportunity to apply what they had learned in the classroom.

Analyses

The data from the TPISD were analyzed under the assumptions of True Score theory and Item Response theory. Analyses based on both theories were used to investigate how each measurement model produced evidence for stability of the TPISD over time and produced evidence for change in teachers' perceptions.

Stability Analyses

True Score Theory Analyses. Temporal stability of a measure addresses how constant scores remain from one occasion to another (Devellis, 1991). A two-score method of computing reliability was conducted using a coefficient known as a Pearson product-moment correlation coefficient which correlated total scores from both administrations. In addition, results of the factor and item analysis for the second administration of the survey were reviewed to further explore the stability of these results.

Item Response Theory Analyses. The analyses for determining stability of the TPISD instrument followed the steps outlined by Wright (1996) with the use of the WINSTEPS computer program (Wright & Linacre, 1998). This method was chosen (over the use of the FACETS model where time is an added facet) because it includes a correction procedure for item and step calibrations found to be variant over time. The method used in this study began by pairing estimates (calibrations) for each person p, item d and rating scale step f. Rating scale calibrations were obtained for the total scale rather than allowing them to be unique for each item. The item calibrations (d) were first

plotted on to an XY graph for a visual picture of the comparison. Standardized differences were then computed between each pair of item and rating scale step calibrations. The formula for the standardized differences between any pair of parameters is:

$$z = (d_1 - d_2) / (s_1^2 + s_2^2)^{1/2}$$

where s is the standard error of the parameter. The standardized difference between different estimates of the same parameter has an expectation of zero and a variance of one (Wolfe & Chiu, 1999). Values of $|z|$ greater than 2.00 are considered large enough to indicate unstable item calibrations or step calibrations across time periods.

Measurement of Change

One purpose for the development of the TPIDS was to create a standardized instrument that could be utilized for measuring change in teachers' perceptions of the impact that staff development had on their teaching over time. How useful it is for this purpose speaks to its validity. If during the stability analyses it was determined that the identity of the variable did not remain stable over the two occasions, an equating method originally proposed by Wright (1996) and utilized by Wolfe and Chiu (1999) was to be carried out. The purpose of this method is to separate changes in persons from changes in rating scale functioning. The method is based on item response theory for which a counterpart in true score theory does not exist.

Results

Reliability Analyses

True Score Theory Analyses. Stability reliability over the two administration of the TPISD as measured by a Pearson correlation was strong, $r = 84$, $p < .01$. Evaluation for outliers by conducting a linear regression with scores from Time 1 to predict scores at Time 2 produced one case with a standardized residual of 3.39. This case was dropped before the final Pearson correlation was derived. The one-factor structure of the scale remained stable as indicted by a principal components analysis conducted at both time points. Scoring patterns also remained stable as indicated by the analysis of item and scale statistics (means and standard deviations). The derived stability coefficient was of sufficient strength to say that TPISD scores remained stable across two administrations, yet the presence or absence of measured change can be due to other things besides the reliability of an instrument including changes in other facets of the measurement situation such as interpretation of the items or use of the rating scale (Wolfe & Chiu, 1999).

Item Response Theory Analyses. To evaluate the invariance of item and step calibrations, the item calibrations were first compared for the set of 25 items across the two administrations. A plot of the item calibrations from the two administrations of the TPISD is presented in Figure 1.

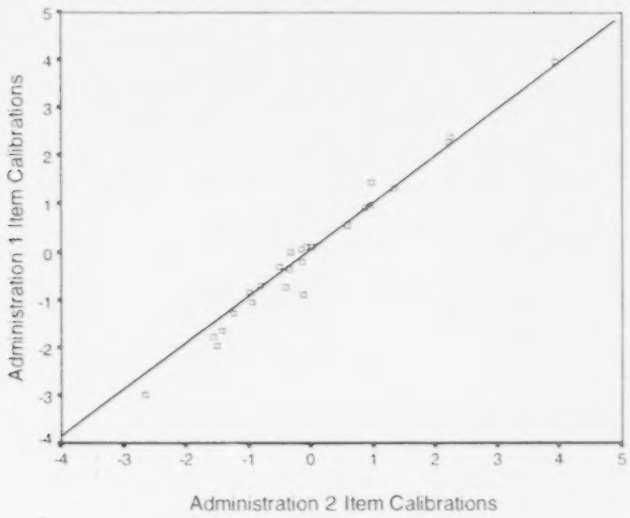

Figure 1. Item calibrations from Time 1 plotted against item calibrations from Time 2.

Visual inspection of Figure 1 shows that most items fall close to the identity line. One item near the center of the plot appears to fall away from the line more than other items, and is a flag that at least one item logit will be found to vary significantly between administration one and two.

After visually inspecting for invariant items, standardized differences between item calibrations and step calibrations were calculated by using the formula:

$$z = (d_1 - d_2) / (s_1^2 + s_2^2)^{1/2}$$

where s is the standard error of the parameter. The values for the derived standardized differences between item calibrations are presented in Table 1.

Table 1

Standardized Values for Item Calibrations

Item	z	Item	z	Item	z
1	-1.02	10	1.14	18	.77
2	-.73	11	.00	19	-.24
3	-1.74	12	.86	20	.05
4	.57	13	.51	21	1.6
5	-.39	14	-1.37	22	2.39
6	.91	15	-3.42	23	.91
7	.29	16	.19	24	.52
8	.25	17	-.04	25	-.05
9	-.87				

The standardized difference values revealed two items, 15 and 22, with values outside the range of $|z| > 2.00$. At a 95% confidence level, we would expect only one value to be outside this range by chance.

Standardized difference values were calculated for the step calibrations using the same procedure as for the item calibrations. These values are presented in Table 2 along with the step calibrations and standard errors.

Table 2

Step Calibrations and Standardized Differences

Scale Step	Time 1 Calibration	Time 1 SE[a]	Time 2 Calibration	Time 2 SE	z
1 to 2	-.311	.11	-2.77	.11	-2.19
2 to 3	.08	.05	-.14	.06	2.82
3 to 4	3.03	.04	2.91	.05	1.87

Standardized differences for the step calibrations revealed that two of three rating scale steps were used differently at administration one and two. These statistics combined with the standardized difference values for the items suggest that interpretation of change in impact on teaching as reflected by differences in TPISD total scores from

administration one and two would be confounded with change in the use of the rating scale.

With the noted variance of some items and step calibrations, the method developed by Wright (1996) and demonstrated by Wolfe and Chiu (1999) was utilized to correct for this variance. The method is a five step process, with step one being the derivation of standardized difference scores to determine if variance of item and step measures exists. The second step involves correcting the step calibrations so a common rating scale for the two administrations is created. To do this, the data set from administration one and two were stacked to form one data set (each person in administration two was given a different identification number). The stacked data set had 305 persons' responses to the 25 TPISD items, being comprised of the responses from two surveys that 152 persons completed. The basic rating scale analysis was then repeated and a new set of step calibrations was obtained. All other values obtained such as item and person measures were ignored. The values for the corrected step calibrations are presented in Table 3. The next steps in the analysis used the corrected step calibrations values.

Table 3

Uncorrected and Corrected Step Calibrations

Scale Step	Time 1 Calibration	Time 2 Calibration	Corrected Calibration	Standard Error
1 to 2	-.311	-2.77	-2.93	.08
2 to 3	.08	-.14	-.03	.04
3 to 4	3.03	2.91	2.95	.03

In the third step of the analysis, corrected person and item calibrations were obtained for administration one data by anchoring rating scale steps on the values obtained above. Anchoring was done by using the data from administration one and running another

rating scale analysis, but the rating scale step calibrations were *forced to be the corrected* calibrations listed in Table 3. The corrected item and person calibrations for administration one were then used in the last two steps in the analysis as the basis for measuring change in person and item calibrations over the two administrations.

For step four, the administration two data were re-analyzed by anchoring the step calibrations on the common-scale values obtained from the stacked data set during step two. In addition, the twenty-three invariant or stable items from the initial analyses were anchored on the corrected item calibrations from step three. Those items that were not invariant, items 15 and 22, were not anchored. From this analysis, new person measures were obtained which were considered to be corrected administration two measures that are referenced to a rating scale that is valid for both administration one and administration two. In addition, the item calibrations were now considered to be a set of item calibrations that are invariant across time. Because the person measures had been corrected for the variance in item and step calibrations, change was attributed to true change in perceptions rather than change in the interpretations of items or use in the rating scale over time.

This change in teacher perceptions was then controlled for in the fifth step of the analyses. Here, the administration two data were re-calibrated by anchoring the scale steps on the joint calibrations obtained from step two, and anchoring the person measures on the corrected estimates from step four. All the items however, were allowed to float (were not anchored). This resulted in item calibrations for each item at administration two that were corrected for changes in both the interpretation of the rating scale and person changes over time.

Results at step five revealed that item 22 no longer had a standardized difference greater than $\pm$ 2.00. This meant that what appeared to be a change or variance in how item 22 was perceived was actually an artifact due to changes in teachers and/or changes in interpretation of the rating scale. This left item 15 as the only item that varied significantly across time.

Measurement of Change

True Score Theory Analyses. In order to examine change in teachers' perceptions over time, data from administration one and two were first analyzed to determine if the assumptions of ANOVA were met. The intent was to run a repeated measures ANOVA with type of staff development as a between subjects factor. For raw score data, both Box's M and Levene's Test produced statistics indicating a violation of the homogeneity of variance assumption. Review of variances for each group across the two time periods found that many of the larger variances were paired with the smaller groups which creates a positive bias in the *F* statistic used in the significance test (Keppel, 1991). Given the violation of the homogeneity of variance assumption along with the sharply unequal group sizes ranging from 2 to 29, it was decided that a repeated measures ANOVA using raw score data would not be appropriate.

Instead, a paired sample t-test was first conducted with the overall means on the TPISD for administration one and two. The result of this analysis was not significant, $t = 1.104$, $p = .272$. Paired sample t-tests run for each staff development group using a Bonferroni correction also revealed no significant differences.

For the person logit measures obtained using IRT, evaluation of ANOVA assumptions found no violation of the homogeneity of variance assumption. Both the

Box's M and Levene's Test statistics were nonsignificant. A repeated measures ANOVA with type of staff development as a between subjects factor was found to be nonsignificant, $F = .010$, $p = .920$.

Item Response Theory Analyses. The corrected administration one and two person measures from the stability analysis (steps three and four) were used to further investigate how much change in teachers perceptions occurred across administrations. Because the person measures had been corrected for the variance in item and step calibrations, change was attributed to true change in perceptions rather than change in the interpretations of items or use in the rating scale over time.

Of the 152 teachers, 26.3% (n=40) reported a significant change in their perceptions of the impact that a staff development had on their teaching as measured by standardized difference scores greater than $\pm$ 2. One-half of those teachers (n=20) reported significantly more impact over time and the other half (n=20) reported significantly less impact over time. Before the person measures were corrected, the pattern of results were similar but not equal. With the uncorrected person measures, 14 teachers (9%) reported significantly more impact over time and 26 teachers (17%) reported significantly less impact over time. By correcting the person measures for the variance of items and steps, conclusions about 8% (n=12) of the teachers were changed. Overall, the correction for variance provided a slight negative shift in z values which translated into teachers reporting more impact of staff development on their teaching across time

Discussion

The findings from this study addressed the stability of the TPISD over two administrations. The information gained for this inquiry is particularly important if the TPISD is to be used for comparing teacher perceptions over time. True Score theory results of correlating administration one and two total scores revealed a strong Pearson correlation of .84. The Pearson correlation of corrected person measures obtained from the item response theory analysis closely matches this result (.82).

How does one determine whether any lack of stability in scores across time periods is due to the instability of the construct, the instability of the instrument, or change in the reported amount of the latent variable over time? As noted by DeVellis (1991), the examination of change in scores over time (using true score theory) should be thought of as an investigation into "temporal stability" where change can be the result of a variety of things besides the reliability of the instrument. The evidence for stability in this study gathered from the True Score analysis (Pearson correlation) was thus thought to be a combination of evidence for measurement stability, construct stability and change in the level of the construct reported by teachers over time.

Item Response theory has an advantage over True Score theory when evaluating temporal stability because it allows for the examination of measurement stability apart from changes in the level of the construct demonstrated by persons over time. The method used for this differentiation is possible because Item Response theory derives a standard error for each individual measure of items, steps, and persons and thus, standardized differences for all measures can be calculated. The evaluation of variance in

items and step measures in the present study revealed two items and two steps which varied significantly between the two time periods. Utilizing a correction procedure, one of the items and both of the step measures were recalibrated and no longer found to be variant. By evaluating the TPISD with this methodology, more certainty about the stability of the scale was gained than with the use of true score theory.

One item during the Item Response theory analysis was found to be variant across time periods despite the utilization of a correction procedure. This item read, "I am collaborating with other teachers on the use of this innovation." Calibrations for this item indicated it become significantly more difficult to agree with over the four month period. What could have led to this shift in item difficulty? Smith (1996) suggests that the significant shift in logit values not be directly interpreted as an indication of an unstable item, but rather, that the analysis of response frequencies be conducted to further investigate what might have caused this shift in value. A review of responses to the variant item found a shift downward in the number of teachers agreeing with this statement. A closer analysis revealed that 23% of the teachers who originally answered "*Strongly Agree*" actually had missing values on the second survey. Had these teachers responded to the question, perhaps the response category percentages would have been more stable and so perhaps would have the item calibration. On the other hand, 66% of the teachers who changed their response to this item shifted from "*Strongly agree*" to "*Agree.*" This would indicate that either their interpretation of the item had changed, or it was truly more difficult to collaborate with other teachers as the school year went on. The later explanation is certainly plausible but neither explanation can be substantiated without further data.

Another part of this research study addressed change in scores over time and asked what evidence of change in scores was provided by True Score theory versus Item Response theory. To evaluate change in scores, True Score theory is limited to group level comparisons. In the present study, this evaluation of change in scores across groups was further limited by an unexpected change in the number of teachers reporting on different staff developments. More staff developments than expected had actually been taken creating highly unequal *n*'s across groups. Despite this design problem, a True Score theory comparison using IRT person logit measures across teachers was possible and did not reveal a significant change in scores over the four month time period.

Was there any change in scores over time? Results of the change analysis utilizing Item Response theory seemed to provided an answer to this question. Because the analysis using IRT provided individual error terms for each person's score, standardized differences could be computed and these were evaluated to determine if a significant change in *individual* scores had occurred over time. Results of this analysis revealed that 26% ($n = 40$) of the sampled teachers had a significant change in scores, with one-half reporting more impact and one-half reporting less. Further analysis was then possible to see if a significant number of these teachers had taken the same staff development or if the changed scores were dispersed randomly among groups. Of those teachers reporting less impact over time, 30% (n=6) were noted to have taken the same staff development.

Another notable finding was that 30% (n=6) of the teachers who reported significantly more impact from a staff development over time were from one school. For the overall sample, this particular school represented just 7% (n=11) of the teachers in the study. These six teachers took five different staff developments, so factors other than

one highly impacting staff development would seem to be responsible for the reported increase in impact over time.

In summary, use of True Score theory for the evaluation of change in scores on the TPISD across a four month time period provided no evidence that any change had occurred. Item Response theory analysis gave evidence that indeed, 26% of the scores changed significantly across the four month time period which is far more than would be expected by chance using a 95% confidence interval. In addition, by using corrected person scores obtained during the stability analysis, more certainty was gained that the change in scores was due to changes in the level of the measured variable rather than changes in the measure itself.

Collins, L. M. (1996). Is reliability obsolete? A commentary on "Are simple gain scores obsolete". Applied Psychological Measurement, 20(3), 289-292.

Devellis, R. F. (1991). Scale Development. Newbury Park, CA: SAGE.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). Educational research: An introduction. (6th Ed. ed.). White Plains, N.Y.: Longman.

Keppel, G. (1991). Design and analysis: A researcher's handbook. (3rd Ed. ed.). Englewood Cliffs, N.J.: Prentice-Hall, Inc.

Linacre, J. M. (1998, April). Investigating Rating Scale Category Utility . Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Smith, R. M. (1996). Pre/post comparisons in Rasch measurement. In G. Engelhard & M. Wilson (Eds.), Objective Measurement: Theory into Practice, Volume 3 . Norwood, NJ: Ablex.

Williams, R. H., & Zimmerman, D. W. (1996). Commentary on the commentaries of Collins and Humphreys. Applied Psychological Measurement, 20(3), 295-297.

Wolfe, E., W., & Chiu, C. W. T. (1999). Measuring pretest-posttest change with a Rasch rating scale model. Journal of Outcome Measurement, 3(2), 134-161.

Wright, B. D. (1996). Time 1 to time 2 comparison. Rasch Measurement Transactions, 10(1), 478-479.

Wright, B. D., & Linacre, J. M. (1998). WINSTEPS (version 2.82) [Computer software]. Chicago: MESA Press.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, Illinois: Mesa Press.